An EM Algorithm for Independent Component Analysis in the Presence of Gaussian Noise

LETTER

An EM Algorithm for Independent Component Analysis in the Presence of Gaussian Noise

Mingjun Zhong and Huanwen Tang
Institute of Computational Biology and Bioinformatics, Dalian University of Technology
Dalian 116023, People's Republic of China
E-mail:sunxl_zhong@yahoo.com

Huili Wang
Institute of Neuroinformatics, and Department of Foreign Languages
Dalian University of Technology, Dalian 116023, People's Republic of China
E-mail:graduate@dlut.edu.cn

Yiyuan Tang
Institute of Neuroinformatics, Dalian University of Technology
Dalian 116023, People's Republic of China
Laboratory of Visual Information Processing and Key Lab for Mental Health
The Chinese Academy of Sciences, Beijing 100101, People's Republic of China
E-mail:yy2100@163.net

(Submitted on November14,2003)

Abstract—An expectation-maximization (EM) algorithm for independent component analysis in the presence of gaussian noise is presented. The estimation of the conditional moments of the source posterior can be accomplished by maximum a posteriori estimation. The approximate conditional moments enable the development of an EM algorithm for inferring the most probable sources and learning the parameter in noisy independent component analysis. Simulation results show that the proposed method can perform blind source separation of sub-Gaussian mixtures and super-Gaussian mixtures.

Keywords—Independent component analysis, blind source separation, EM algorithm, maximum a posteriori

1. Introduction

Independent component analysis (ICA) [1, 2, 3, 4, 5] is a statistical technique whose goal is to recover independent components given only observations that are unknown linear mixtures of the unobserved independent source signals. One may formulate ICA as the estimation of the following linear generative model for the data:

$$x = As + \varepsilon, \tag{1}$$

where $x = (x_1, \cdots, x_N)^T$ is the vector of observed random variables, $s = (s_1, \cdots, s_M)^T$ is the vector of the independent components, $A \in R^{N \times M}$ is the unknown constant mixing matrix, and the vector $\varepsilon = (\varepsilon_1, \cdots, \varepsilon_N)^T$ is the noise modelled as Gaussian with zero mean and covariance matrix Σ. In the expectation-maximization (EM) algorithm [6] proposed in this paper, we assume the dimension of s equals the dimension of x, i.e., $N = M$, and the noise covariance matrix Σ is known.

It is shown in this paper that the conditional moments of the source posterior distribution can be estimated by maximum a posteriori (MAP) estimation. Given these approximate conditional moments, an EM algorithm is presented for learning the parameter A in equation (1). To illustrate the EM algorithm proposed in this paper, simulation results are presented for blind source separation (BSS) [2, 7] of super-Gaussian mixture and sub-Gaussian mixture signals.

2. The Approximate Conditional Moments

The approximate conditional moments of the posterior can be derived by modelling the posterior as gaussian. Assuming that we have observed T data samples $x = x(1), \cdots, x(T)$ which are generated according to model (1), the approximate conditional moments for the gaussian posterior distribution, based on the Laplace estimation [8, 9], are given as (see appendix A):

$$E\{s|x(t)\} = \hat{s} = \arg\max_{s} L(s) \tag{2}$$

$$E\{ss^T|x(t)\} = H(\hat{s})^{-1} + \hat{s}\hat{s}^T, \tag{3}$$

where $L(s) = log\{p(s|x(t), A)\}$, where p denotes the probability distribution function (p.d.f.) which is used throughout this paper, is the log-posterior, and the Hessian of the approximate log-posterior computed at the MAP value $\hat{s}$ is denoted as $H(\hat{s}) = -\nabla_s \nabla_s log\{p(s|x(t), A)\} = A^T \Sigma^{-1} A - \nabla_s \nabla_s log\{p(\hat{s})\}$ where ∇_s denotes the gradient with respect to s used throughout this paper.

The estimation of the independent components can be accomplished by MAP estimation. According to the generative model (1), the log-posterior, i.e., $L(s)$, is given as follows:

$$L(s(1), \cdots, s(T)) = \sum_{t=1}^{T} \{-\frac{1}{2}(x(t) - As(t))^T \Sigma^{-1}(x(t) - As(t)) + \varphi(s(t))\} + C, \tag{4}$$

where $\varphi(s(t)) = log\{p(s(t))\}$ and C is a constant irrelevant to $s(t)$. To estimate the independent components, taking the gradient of the log-posterior with respect to $s(t)$ gives the following gradient learning rule:

$$\nabla_s L(s(t)) = A^T \Sigma^{-1}(x(t) - As(t)) + \varphi'(s(t)), \tag{5}$$

where $\varphi'(s(t)) = \nabla_s \varphi(s(t))$ is called activation function in some literatures [7, 10]. Note that the log-posterior in equation (4) is essentially the same as the likelihood proposed by Hyv$\ddot{a}$rinen [3]. It is indicated in the following section that the approximate conditional moments in equations (2) and (3) enable the development of an EM algorithm for learning the parameter A in model (1).

3. An EM Algorithm for Parameter Estimation

To derive a learning algorithm for inferring A in model (1), it is required to maximize the following data likelihood given the model:

$$p(x|A) = \int p(x|s, A)p(s)ds. \tag{6}$$

Rather than using some approximations to this intractable integral as in [8, 11], it is desirous to employ the EM framework for estimation and inference for this form of linear model, as this is a most natural method for maximizing the data likelihood. Given the approximate conditional moments, the standard form of M-step for A emerges (see appendix B):

$$A^{new} = (\sum_{t=1}^{T} x(t)E\{s|x(t)\}^T)(\sum_{t=1}^{T} E\{ss^T|x(t)\})^{-1}. \tag{7}$$

Inserting equations (2) and (3) into (7) and setting $\Lambda(t) = -\nabla_s \nabla_s \varphi(\hat{s}(t))$, we have:

$$A^{new} = (\sum_{t=1}^{T} x(t)\hat{s}(t)^T)[\sum_{t=1}^{T} \{(A^T \Sigma^{-1} A + \Lambda(t))^{-1} + \hat{s}(t)\hat{s}(t)^T\}]^{-1}. \tag{8}$$

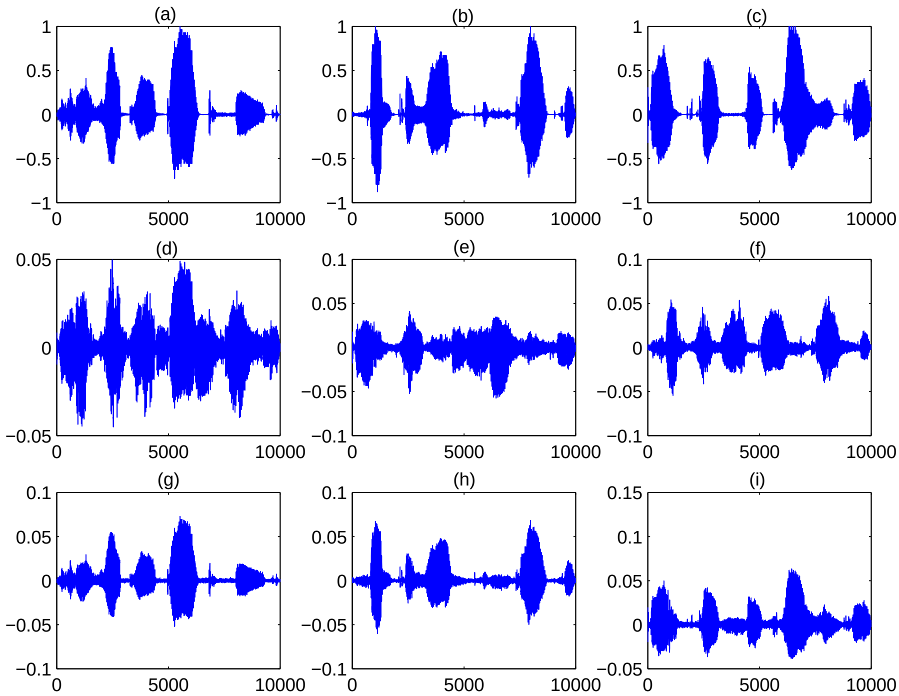

Figure 1. (a-c)The original sources.(d-f)The noisy observations.(g-i)The inferred sources using the EM algorithm outlined.

Since $\hat{s}(t)$ must be inferred in each M-step, a simple alternating variable method, which has already been used in similar estimation tasks [3], should be derived for learning and inference. The method is based on first the optimization of the objective function with respect to $s(t)$ for fixed A, then the optimization with respect to A for fixed $s(t)$, and so on. The optimization with respect to $s(t)$ for fixed A is accomplished by equation (5), and the optimization with respect to A for fixed $s(t)$ is accomplished by equation (8). Thus, the EM procedure of the following form is obtained:
(i)Taking some initial value for A^0. Set $s^0 = (A^0)^+ x$ where $(A^0)^+$ denotes the Moore-Penrose pseudo-inverse of A^0, and let $k = 1$. Normalize each column of A^0 to have unit norm.
(ii)Set $s^k(t) = s^{k-1}(t) + \eta \nabla_s L(s^{k-1}(t))$, using A^{k-1} as the estimate of A.
(iii)Compute A^k by equation (8), using A^{k-1} as the estimate of A and substituting $s^k(t)$ for $\hat{s}(t)$. Normalize each column of A^k to have unit norm.
(iv)Set $k = k + 1$, and go back to step (ii) if not converged.
In this algorithm η refers to some learning rate.

4. Simulations

We applied the EM algorithm of our paper for blind separation of independent components from noisy mixtures. According to the dichotomy proposed by Hyvärinen [3], to illustrate the EM algorithm proposed in this paper, it would be enough to use just two different densities in equation (4), one corresponds to a distribution that is super-Gaussian (i.e., has a positive kurtosis) and another one corresponds to a distribution that is sub-Gaussian (i.e., has a negative kurtosis).

For the inference and estimation in the case of super-Gaussian independent components, three speech signals* are used. Each speech signal has a positive kurtosis. The activation function $\varphi'(s) = -tanh(s)$ [2, 7, 10] is employed in this simulation. The three speech signals were mixed by a randomly chosen matrix A, and noisy data with zero mean and covariance matrix $\Sigma = 0.01I$ were added to the mixtures. The EM algorithm proposed was used for learning A and inferring the independent components s. The proposed algorithm procedure was randomly

*Where the three speech signals were obtained from http://www.cnl.salk.edu/~ tewon/Over.

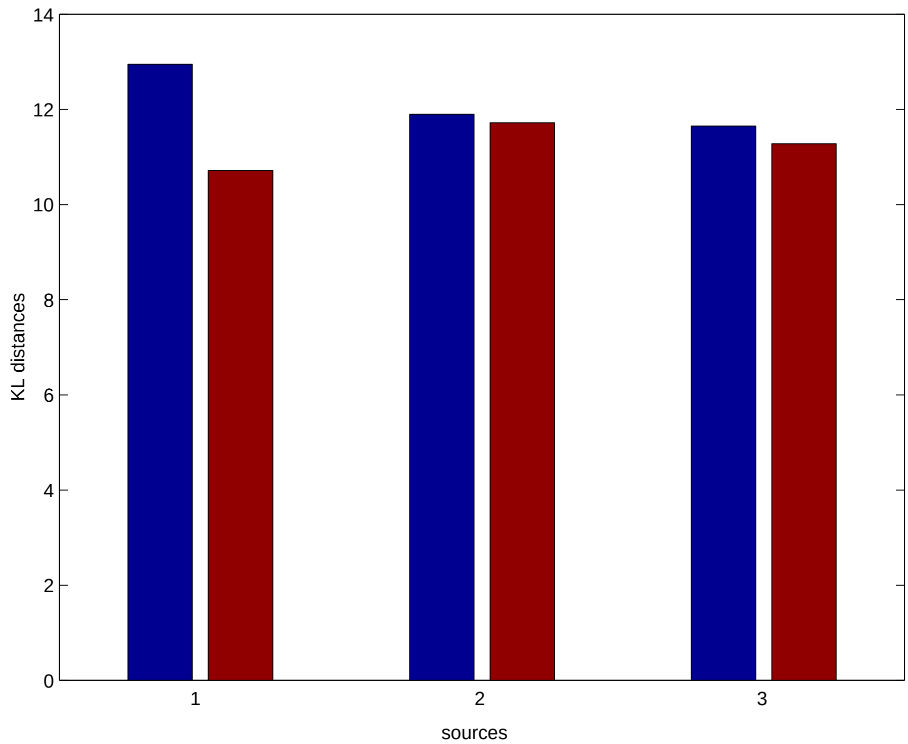

Figure 2. The KL distances between the super-Gaussian sources and the corresponding recovered sources. The left and right bars in each group represent the absolute values of the logarithms of the KL distances between the sources and the recovered sources obtained by the EM algorithm and the method reported in [3], respectively.

initialized and converged in about 20 parameter updates. Figure 1 shows 10,000 samples of the original sources (a-c), the noisy observations (d-f) and the inferred sources (g-i). To assess the accuracy of the recovered sources, we compute the Kullback-Leibler (KL) distance [12] between the source density and the recovered source density. Since both densities are assumed to be super-Gaussian, the KL distance can be estimated via [13]:

$$K_m = \int p(s_m) log \frac{p(s_m)}{p(\hat{s}_m)} ds_m \approx \sum_{t=1}^{T} log \frac{p(s_m(t))}{p(\hat{s}_m(t))}, \quad (9)$$

where $p(s) \propto cosh^{-1}(s)$, and $m = 1, \cdots, M$. For the comparative purpose, we also computed the KL distances between the source densities and the recovered source densities obtained by the method reported in [3] under the same requirements. Figure 2 shows the absolute values of the logarithms of the KL distances. Note that since the logarithms of the KL distances are negative values, the KL distances decrease respectively over those obtained by the method reported in [3].

In the case of sub-Gaussian independent components, the cubic nonlinear function, i.e., $\varphi'(s) = -s^3$, has been a favorite choice as the activation function in equation (5) [7]. Binary and uniform distributions are used as the original distributions of the independent components. Each original signal shown in Figure 3 has a negative kurtosis. The “sign” nonlinearity which was originally used in [3] is employed for estimating A. Hence, the M-step in equation (8) has a new form of $A^{new} = (\sum_{t=1}^{T} x(t) sign(\hat{s}(t)^T))[\sum_{t=1}^{T}\{(A^T\Sigma^{-1}A + \Lambda(t))^{-1} + sign(\hat{s}(t)) sign(\hat{s}(t)^T)\}]^{-1}$, where the “sign” operator is applied separately on each component of its argument in this simulation. The three signals were mixed by a randomly chosen matrix A, and noisy data with zero mean and covariance matrix $\Sigma = 0.01I$ were added to the mixtures. Figure 3 shows the original sources (a-c), the noisy observations (d-f), the inferred sources without using the “sign” nonlinearity (g-i) and the inferred sources using the “sign” nonlinearity (j-l). Figure 4 shows the absolute values of the logarithms of the KL distances. Note that since the logarithms of the KL distances are negative values, the KL distances decrease respectively over those obtained by the method reported in [3].

These two simulations serve to show that the EM algorithm proposed in this paper is able to estimate the

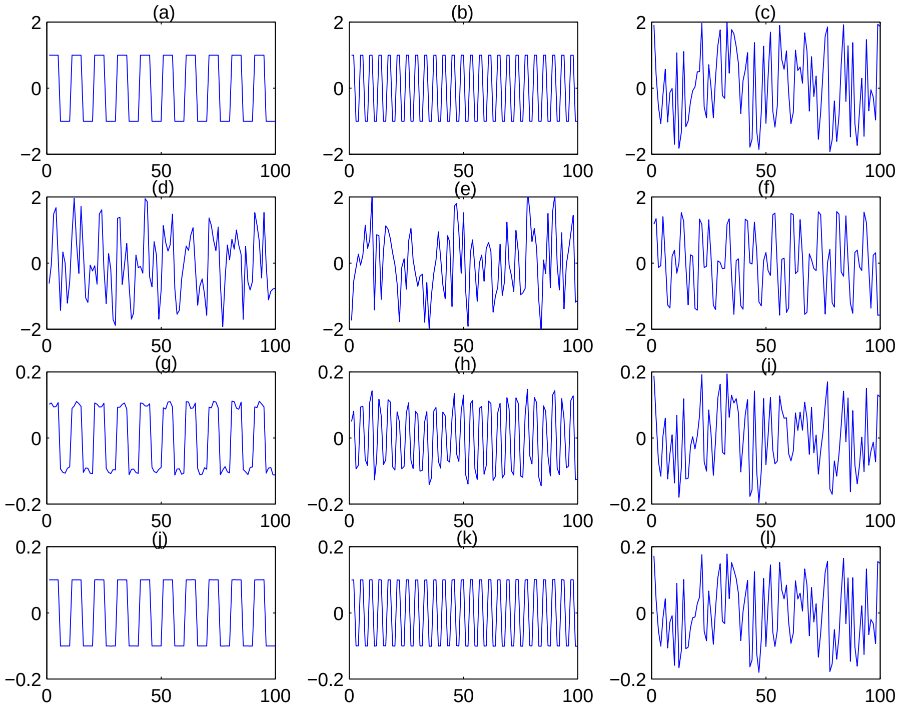

Figure 3. (a-c)The original sources.(d-f)The noisy observations.(g-i)The inferred sources without using the "sign" nonlinearity. (j-l)The inferred sources using the "sign" nonlinearity.

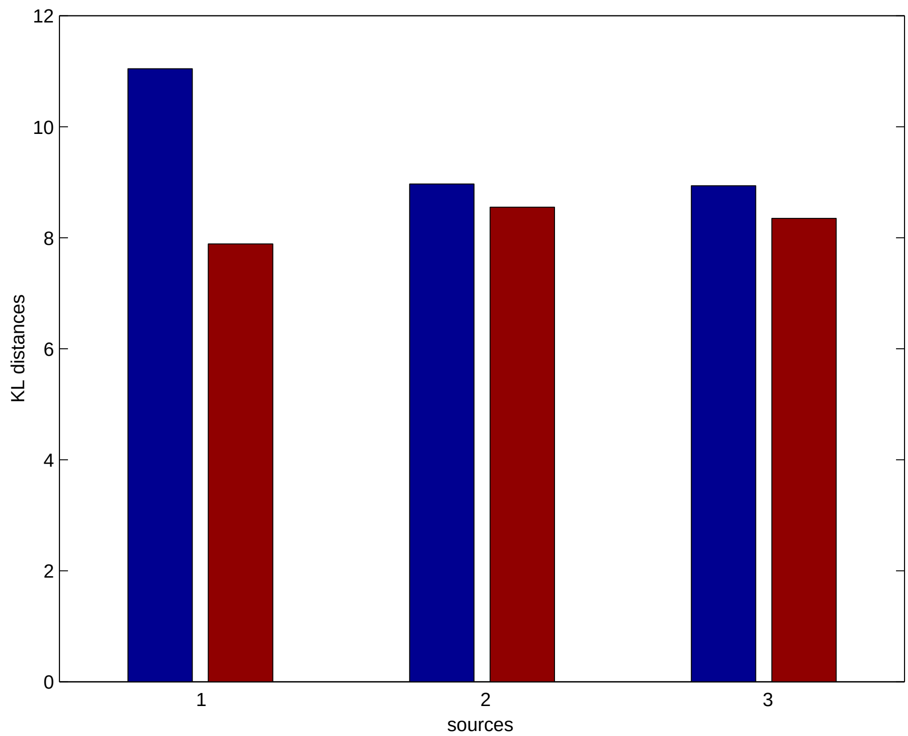

Figure 4. The KL distances between the sub-Gaussian sources and the corresponding recovered sources. The left and right bars in each group represent the absolute values of the logarithms of the KL distances between the sources and the recovered sources obtained by the EM algorithm and the method reported in [3], respectively.

mixing matrix and infer the most probable independent components in noisy ICA.

5. Conclusions

We have proposed an EM algorithm for learning the parameter A and inferring the most probable independent components in noisy ICA. The approximate conditional moments of the posterior distribution estimated by MAP estimation are used to develop an EM algorithm for learning the parameter and inferring the most probable independent components.

Some work on ICA in the presence of Gaussian noise has been done in some literatures [3, 10]. One method more related to ours was introduced in [13], in which the EM algorithm was used for independent factor analysis. The proposed method is simple and computationally less demanding than the one proposed in [13]. In particular, the complexity of the method proposed in [13] is exponential as a function of the number of the independent components, and the algorithm in this paper has polynomial complexity. Future work will estimate the noise covariance matrix in this EM framework.

Appendix A. Derivation of the Approximate Conditional Moments

Based on the Laplace estimation [8], the posterior distribution of the sources can be modelled as a gaussian distribution:

$$p(s|x(t)) \approx (2\pi)^{-\frac{M}{2}}|H(\hat{s})|^{\frac{1}{2}}exp\{-\frac{1}{2}(s-\hat{s})^T H(\hat{s})(s-\hat{s})\}, \tag{10}$$

where $\hat{s}$ is the MAP value and $H(\hat{s})$ is the Hessian of the approximate log-posterior computed at the MAP value. Thus, the approximate conditional moments in equations (2) and (3) can be derived by this gaussian posterior distribution.

Appendix B. Derivation of the M-step

To derive the M-step for the parameter A, it is required to maximize the expected value of the complete data likelihood of the following form given the observed data x and the current model [6]:

$$Q(A^{new}|A^{old}) = \sum_{t=1}^{T}\int p(s|x(t), A^{old})log\{p(x(t), s|A^{new})\}ds, \tag{11}$$

where A^{old} is the parameter obtained in the previous iteration. Setting the gradient of $Q(A^{new}|A^{old})$ with respect to A^{new} to zero gives the new value of the parameter of the M-step in equation (7).

Acknowledgments: The work was supported by NSFC(30170321,90103033), MOST(2001CCA00700) and MOE (KP0302).

References

[1] P.Comon, “Independent Component Analysis - A New Concept?”Signal Processing, Vol. 36, No. 3, pp. 287-314, April 1994.

[2] A.J.Bell,T.J.Sejnowski, “An Information-maximization Approach to Blind Separation and Blind Deconvolution,”Neural Computation, Vol. 7, No. 6, pp. 1129-1159, Nov. 1995.

[3] A.Hyvärinen, “Independent Component Analysis in the Presence of Gaussian Noise by Maximizing Joint Likelihood,”Neurocomputing, Vol. 22, No. 1-3, pp. 49-67, Nov. 1998.

[4] T-W.Lee,M.Girolami,T.J.Sejnowski, "Independent Component Analysis Using an Extended Infomax Algorithm for Mixed Subgaussian and Supergaussian Sources,"Neural Computation, Vol. 11, No. 2, pp. 417-441, Feb. 1999.

[5] A.Hyvärinen, E.Oja, "Independent Component Analysis: Algorithms and Applications,"Neural Networks, Vol. 13, No. 4-5, pp. 411-430, June 2000.

[6] A.P.Dempster,N.M.Laird,D.B.Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm(with discussion),"Journal of the Royal Statistical Society B, Vol. 39, No. 1, pp. 1-38, Nov. 1977.

[7] A.Cichocki,S.I.Amari, *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications*, Wiley, New York, 2002.

[8] M.S.Lewicki,T.J.Sejnowski, "Learning Overcomplete Representations,"Neural Computation, Vol. 12, No. 2, pp. 337-365, Feb. 2000.

[9] M.Girolami, "A Variational Method for Learning Sparse and Overcomplete Representations,"Neural Computation, Vol. 13, No. 11, pp. 2517-2532, Nov. 2001.

[10] A.Cichocki,S.C.Douglas,S.I.Amari, "Robust Technique for Independent Component Analysis with Noisy Data,"Neurocomputing, Vol. 22, No. 1-3, pp. 113-129, Nov. 1998.

[11] R.M.Neal, *Bayesian Learning for Neural Networks*, Lecture Notes in Statistics, Vol.118, Springer-Verlag, New York, 1996.

[12] T.M.Cover,J.A.Thomas, *Elements of Information Theory*, Wiley, New York, 1991.

[13] H.Attias, "Independent Factor Analysis,"Neural Computation, Vol. 11, No. 4, pp.803-951, May 1999.

Mingjun Zhong studied mathematics and obtained his Master's degree at the Dalian University of Technology, P.R.China. He is currently studying for a Ph.D. degree at the Institute of Computational Biology and Bioinformatics of the Dalian University of Technology, P.R.China. His research interests include independent component analysis, sparse coding, bioinformatics and neuroinformatics.

Huanwen Tang graduated from the Dalian University of Technology in 1963, and currently a professor of mathematics and management engineering at the Dalian University of Technology. His research interest includes computational models and algorithm of human cognition and neural information coding, bioinformatics and neuroinformatics. (Home page: http://brain.dlut.edu.cn)

Huili Wang graduated from the Shanghai International Studies University in 1989, and currently an assistant professor of school of foreign languages at the Dalian University of Technology. Her research interest includes psychal linguistics, neurolinguistics and neuroinformatics.

Yiyuan Tang graduated from the Jinlin University in 1987, and currently a professor of neuroinformatics and neuroscience at the Dalian University of Technology. His research interest includes neuroimaging, cognition and emotion interaction, neuroinformatics. (Home page: http://brain.dlut.edu.cn)